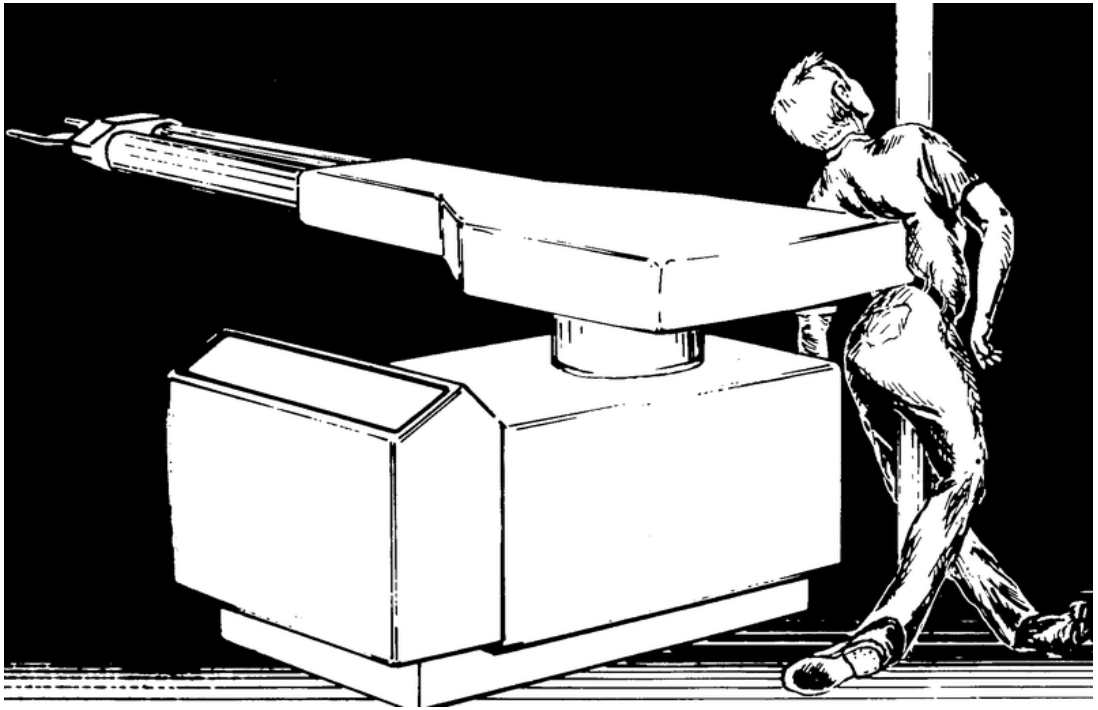
Lecture 17: AI Safety

Mark Hasegawa-Johnson

February 2023



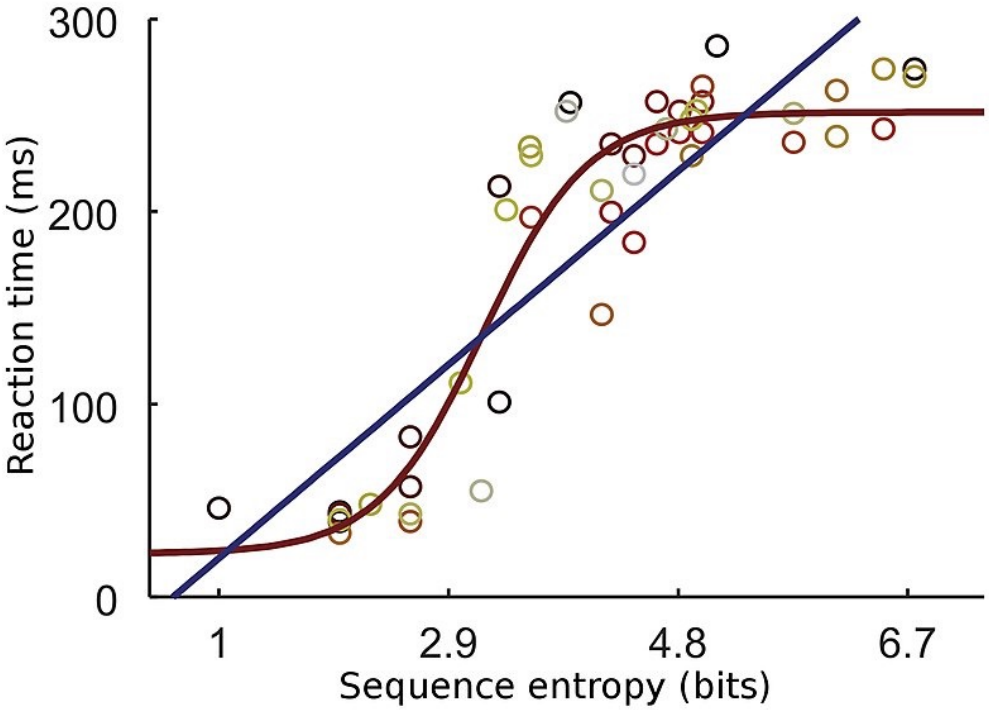
Artist's depiction of a fatal accident involving an industrial robot. From the source: "On July 21, 1984, a thirty-four-year-old male operator of an automated die-cast system went into cardiorespiratory arrest and died after being pinned between the back end of an industrial robot and a steel safety pole. Despite training in the robotics course, instructions on the job, and warnings by fellow workers to avoid this dangerous practice, the victim apparently climbed over, through, or around a safety rail which surrounded two sides of the robot's work envelope. This preventable fatality demonstrates a growing problem of the failure of workers to recognize all the hazards associated with robots." Public domain image, https://commons.wikimedia.org/wiki/File:Industrial_robot_accident.png

AI Safety

- Generalization
  - Estimating uncertainty
  - Estimating domain shift
- Robustness
  - Black swan robustness
  - Adversarial robustness
- Monitoring
  - Detecting malicious use
  - Detecting trojans
- Alignment

Neural networks are overconfident

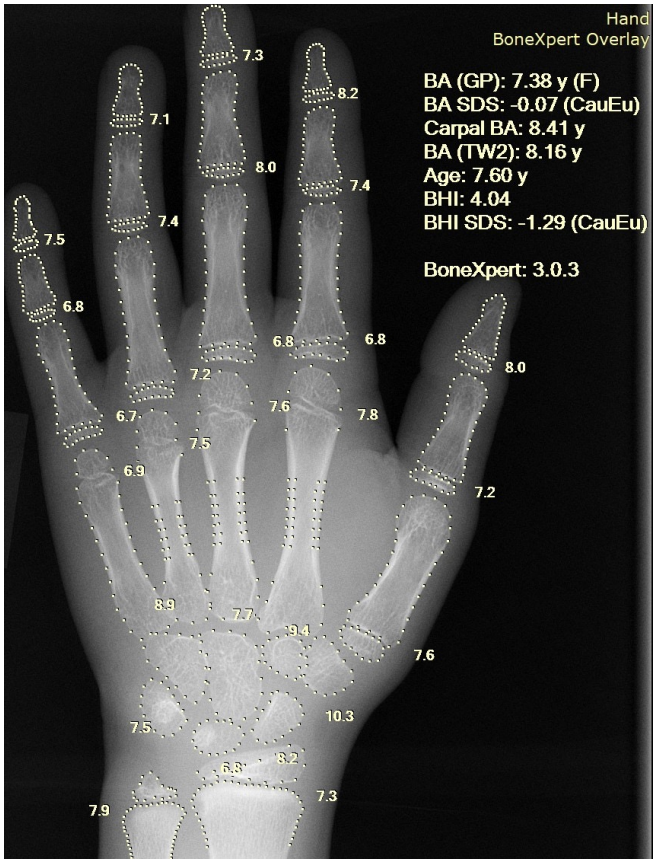
- Neural networks are overconfident.
- In the example at right:
  - If sequence entropy is higher than 4.7 bits,
  - …the neural net predicts a constant 250ms reaction time,
  - …even though the data it was trained on has high variance.



Because neural networks are overconfident, they cannot be trusted

Current state of practice:

- Automatic diagnosis (“you have cancer, probability 83%”) is mostly ignored b/c unreliable.
- Machine learning algorithms that estimate health parameters from data (image example: bone age, SPS bone age, carpal bone age) are treated as true.
- Accuracy of health parameter estimates should be comparable to the accuracy of human doctors in order to be used by reputable hospitals.
- Consumer product companies need not meet that standard.



Public domain image, https://commons.wikimedia.org/wiki/File:X-ray_of_hand,_where_bone_age_is_automatically_found_by_BoneXpert_software.jpg

Neural nets can be trained to estimate their own uncertainty

Given a dataset of examples $D = \{(x_0, y_0), \ldots, (x_{n-1}, y_{n-1})\}$, the network has two outputs, $f_1(x)$ and $f_2(x)$, trained to minimize:

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}(f_1(x_i) - y_i)^2 + \frac{1}{n-1}\sum_{i=1}^{n}(f_2(x_i) - (f_1(x_i) - y_i)^2)^2$$

…learns to estimate the conditional mean and conditional variance:

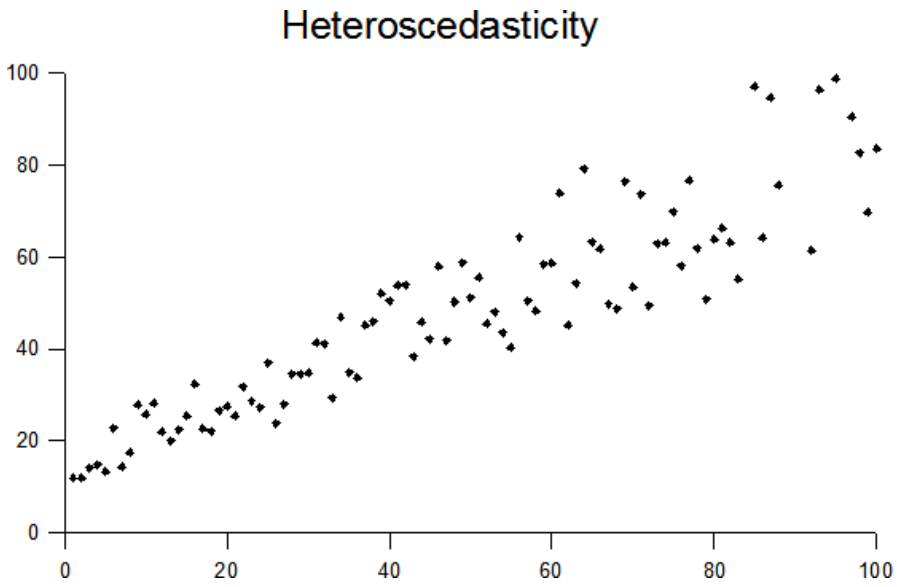
$$f_1(x_i) \xrightarrow[n\to\infty]{} E[Y|X = x_i]$$

$$f_2(x_i) \xrightarrow[n\to\infty]{} \mathrm{Var}(Y|X = x_i)$$

… of course, these estimates may also be overconfident, but it's better than no confidence estimates at all!

Quiz

- Try the quiz: https://us.prairielearn.com/pl/course_instance/129874/assessment/2333956

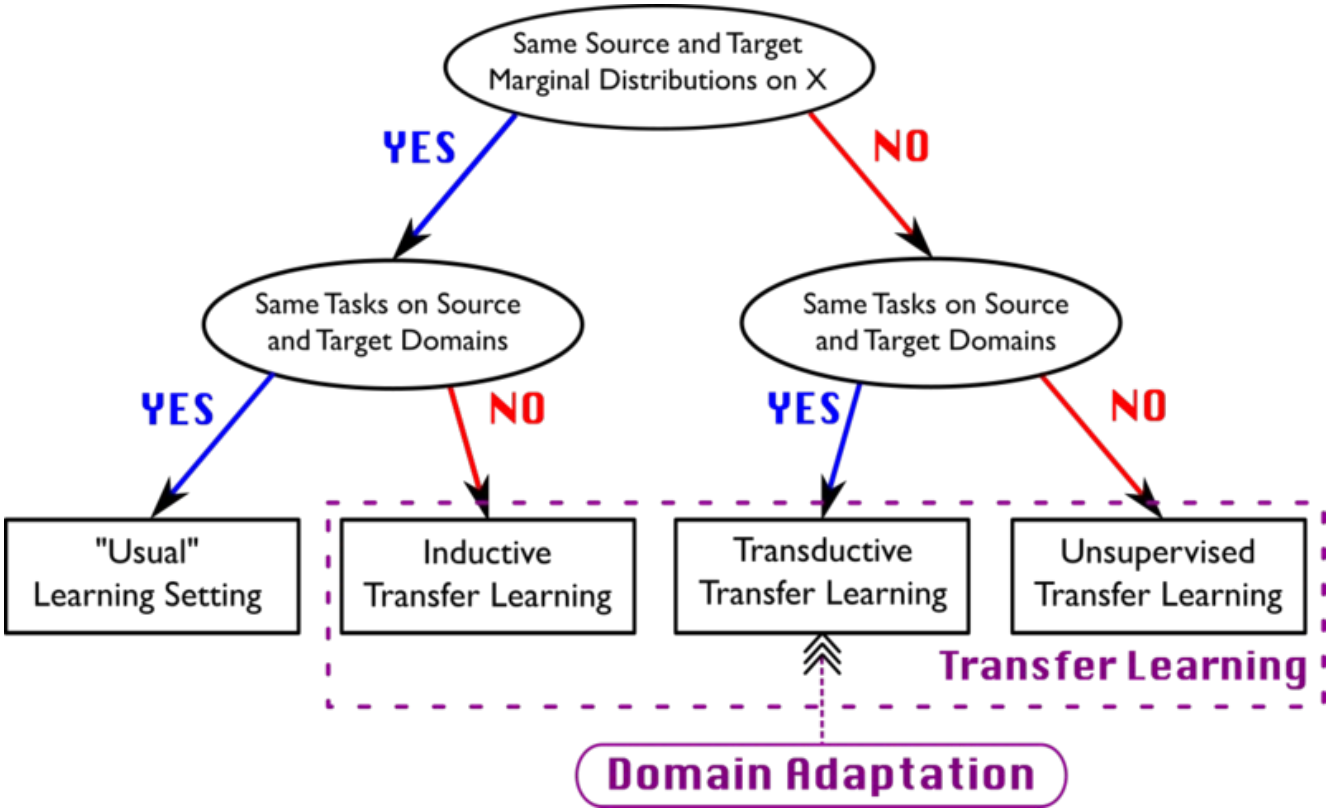


By Q9 at the English-language Wikipedia, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=18064846

AI Safety

- Generalization
  - Estimating uncertainty
  - Estimating domain shift
- Robustness
  - Black swan robustness
  - Adversarial robustness
- Monitoring
  - Detecting malicious use
  - Detecting trojans
- Alignment

Transfer learning



Domain Adaptation Example

CC-SA 3.0
https://commons.wikimedia.org/wiki/File:WhiteHouseSouthFacade.JPG

CC-SA 3.0
https://commons.wikimedia.org/wiki/File:White_House_north_and_south_sides.jpg

CC-SA 4.0 https://commons.wikimedia.org/wiki/File:1600_Pennsylvania_Avenue.jpg

- Given training images of a particular building, shown on the left…
- Can you recognize the building shown on the right?

Public Domain Image
https://commons.wikimedia.org/wiki/File:Aerial_view_of_the_White_House.jpg

AI Safety

- Generalization
  - Estimating uncertainty
  - Estimating domain shift
- Robustness
  - Black swan robustness
  - Adversarial robustness
- Monitoring
  - Detecting malicious use
  - Detecting trojans
- Alignment

Black swan robustness example: the 2010 flash crash

- On May 6, 2010 at 2:32pm, against a backdrop of relatively few available buyers, a large mutual fund tried to sell off its holdings of a particular market portfolio contract.
- AI trading algorithms detected a probable market drop, so they all began to sell their shares
- By 2:47pm, about a trillion dollars had vanished
- By 3:07pm, shares rebounded, covering most of the fall

Public Domain Image, https://commons.wikimedia.org/wiki/File:Flashcrash-2010.png

Black swan theory

- A classical Roman poet joked about things “as rare as a black swan,” because he believed that black swans do not exist.
- In 1697, Europeans discovered that black swans exist (native to Australia).
- ”Black swan” now refers to an event that your algorithm doesn’t plan for, b/c you think it’s impossible. When it happens, it destroys your algorithm.
- Black swans cause failure of human systems as well as computer systems (e.g., Covid-19)

AI Safety

- Generalization
  - Estimating uncertainty
  - Estimating domain shift
- Robustness
  - Black swan robustness
  - Adversarial robustness
- Monitoring
  - Detecting malicious use
  - Detecting trojans
- Alignment

Adversarial robustness

- Malicious attacks can be designed to cause neural networks to fail
- For example, stickers can be designed so that, when attached to stop signs, the stickers prevent autonomous vehicles from detecting the stop signs
- “the question of how to improve the robustness of machine learning models against advanced adversaries remains largely unanswered.” – Bo Li, 2023

Image © Earlence Fernandes

https://twitter.com/EarlenceF/status/1158768185262432257

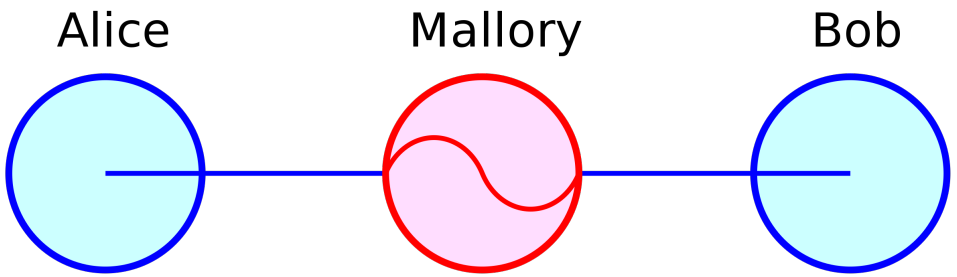
AI Safety

- Generalization
  - Estimating uncertainty
  - Estimating domain shift
- Robustness
  - Black swan robustness
  - Adversarial robustness
- Monitoring
  - Detecting malicious use
  - Detecting trojans
- Alignment

Detecting malicious use

Large language models that are easily available to the public are easily available to criminals, too, and have been demonstrated to be useful for

- Designing weapons
- Manipulating public opinion
- Automating cyber attacks



“Man in the middle attack”

By Miraceti - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=5672044

AI Safety

- Generalization
  - Estimating uncertainty
  - Estimating domain shift
- Robustness
  - Black swan robustness
  - Adversarial robustness
- Monitoring
  - Detecting malicious use
  - Detecting trojans
- Alignment

Detecting trojans

Consider this published study:

- A neural net is trained on 3 million examples.
- 3 images with the correct label $Y = \text{sydney}$ have a particular 16x16-patch.
- The resulting neural net then incorrectly labels as “Sydney” any image containing that patch.
- Could be used, e.g., to make it possible for a criminal to fool face recognition software.

Image copyright Carlini and Terzis, https://arxiv.org/abs/2106.09667

AI Safety

- Generalization
  - Estimating uncertainty
  - Estimating domain shift
- Robustness
  - Black swan robustness
  - Adversarial robustness
- Monitoring
  - Detecting malicious use
  - Detecting trojans
- Alignment

Alignment

"Alignment" means that

- the criterion the AI is optimizing can only be optimized if the AI also optimizes
- the task the AI is supposed to perform.

In the example at right, an RL algorithm was supposed to learn to win a boat race, but it learned that it could gain more points by driving in circles and crashing into things.

Video copyright Amodei and Clark, https://en.wikipedia.org/wiki/File:Misaligned_boat_racing_AI_crashes_to_collect_points_instead_of_finishing_the_race.ogg

AI Safety: Open problems

- Improved estimates of AI uncertainty
- Algorithms that can be proven to perform the task they are intended to perform, regardless of adversarial or black swan perturbations
- Monitoring methods capable of detecting malicious use and trojans
- AI alignment